IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, May 2024

Crime Rate Prediction using Python

Yash Nakhale[1], Chetan Randaye[2], Aditya Sarkar[3], Prof. Pallavi Akulwar[4]
B.TECH Student, Department of Computer Science and Engineering[1,2,3]
Professor, Department of Computer Science and Engineering[4]
Rajiv Gandhi College of Engineering Research and Technology, Chandrapur, India

Abstract: *It has never been easy to invest in a set of assets, the abnormality of the financial market does not allow simple models to predict future asset values with higher accuracy. Machine learning, which consists of making computers perform tasks that normally require human intelligence is currently the dominant trend in scientific research. This article aims to build a model using Machine learning Model. The main objective of this paper is to see in which precision a Machine learning algorithm can predict. Predicting crime prediction rate is a complex task that traditionally involves extensive human-computer interaction.. The network is trained and evaluated for accuracy with various sizes of data, and the results are tabulated. This project is to predict crime to make more acquainted and precise crime in the area*

Keywords: Crime Prediction, Machine Learning, Predictive Policing, Data Analytics, Law Enforcement, Feature Engineering

I. INTRODUCTION

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it to learn for themselves.

The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers to learn automatically without human intervention or assistance and adjust actions accordingly.

But, using the classic algorithms of machine learning, text is considered as a sequence of keywords; instead, an approach based on semantic analysis mimics the human ability to understand the meaning of a text.

Some Machine Learning Methods

Machine learning algorithms are often categorised as supervised or unsupervised.

- Supervised machine learning algorithms can apply what has been learned in the past to new data using labelled examples to predict future events. Starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly.
- In contrast, unsupervised machine learning algorithms are used when the information used to train is neither classified nor labelled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabeled data. The system doesn't figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabeled data.
- Semi-supervised machine learning algorithms fall somewhere in between supervised and unsupervised learning, since they use both labeled and unlabeled data for training – typically a small amount of labeled data and a large amount of unlabeled data. The systems that use this method are able to considerably improve learning accuracy. Usually, semi-supervised learning is chosen when the acquired labeled data requires skilled and relevant resources in order to train it / learn from it. Otherwise, acquiring unlabeled data generally doesn't require additional resources.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, May 2024

- Reinforcement machine learning algorithms is a learning method that interacts with its environment by producing actions and discovers errors or rewards. Trial and error search and delayed reward are the most relevant characteristics of reinforcement learning. This method allows machines and software agents to automatically determine the ideal behavior within a specific context in order to maximize its performance. Simple reward feedback is required for the agent to learn which action is best; this is known as the reinforcement signal.

How to choose the right machine learning model

The process of choosing the right machine learning model to solve a problem can be time consuming if not approached strategically.

Step 1: Align the problem with potential data inputs that should be considered for the solution. This step requires help from data scientists and experts who have a deep understanding of the problem.

Step 2: Collect data, format it and label the data if necessary. This step is typically led by data scientists

Step 3: Choose which algorithm(s) to use and test to see how well they perform. This step is usually carried out by data scientists.

Step 4: Continue to fine tune outputs until they reach an acceptable level of accuracy. This step is usually carried out by data scientists with feedback from experts who have a deep understanding of the problem.

II. LITERATURE SURVEY

In [1] Ling Chen, Xu Lai (2011) has compared the experimental result that is obtained by the ANN (Artificial Neural Network).

In [2] Jyoti Agarwal, Renuka Nagpal, et al., (2013) has studied the crime analysis using K-means clustering on the crime dataset. They have developed this model using the rapid miner tool. The clustered results are obtained and analysed by plotting the values over the years. This model gives the result of the analysis that the number of homicides decreased from 1990 to 2011.

In [3] Shiju Sathyadevan, Devan M. S, et al., (2014) have predicted the regions where there is a high probability of the crime occurred. They have visualized crime-prone areas also. They have classified the data using Naive Bayes classifiers. This algorithm is a supervised learning algorithm that also gives the statistical method for classification. This classification gives an accuracy of the 90%. Lawrence

In [4] McClendon and Natarajan Meghanathan (2015 have used Linear Regression, Additive Regression, and Decision Stump algorithms using the same set of input (features), on the Communities and Crime Dataset. Overall, the linear regression algorithm gave the best results compared to the three selected algorithms.

In [5]Chirag Kansara, Rakhi Gupta, et al., (2016) proposed a model which analyses the sentiments of the people on Twitter and predicts whether they can become a threat to a particular person or society. This model is implemented using the Naive Bayes Classifier which classifies the people by sentiment analysis.

III. METHODOLOGIES

Crime rate prediction involves a blend of statistical analysis, machine learning algorithms, and domain knowledge. Here are some common methodologies used in crime rate prediction:

1. Time Series Analysis: Crime data is often analyzed over time to identify patterns and trends. Time series models like ARIMA (AutoRegressive Integrated Moving Average) or seasonal decomposition methods can be used to forecast future crime rates based on historical data.
2. Regression Analysis: Regression models, such as linear regression or logistic regression, can be employed to understand the relationship between various factors (e.g., socioeconomic indicators, demographic characteristics, policing efforts) and crime rates. These models help identify which factors have the most significant impact on crime rates.
3. Geospatial Analysis: Crime is often spatially clustered, meaning it tends to occur in certain geographic areas. Geospatial analysis techniques, such as spatial autocorrelation analysis, hotspot analysis, and kernel density estimation, can be used to identify high-crime areas and predict future crime rates in those areas.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, May 2024

4. Machine Learning Algorithms: Advanced machine learning algorithms, such as decision trees, random forests, support vector machines, and neural networks, can be applied to crime prediction tasks. These algorithms can handle large and complex datasets and can capture non-linear relationships between predictors and crime rates.
5. Risk Terrain Modeling: This approach focuses on identifying environmental features (e.g., abandoned buildings, liquor stores, public transportation hubs) that are associated with higher crime rates. By mapping these features and analyzing their spatial relationships, researchers can predict where future crimes are likely to occur.
6. Agent-Based Modeling: Agent-based models simulate the behavior of individual actors (e.g., criminals, law enforcement officers, residents) within a given environment. These models can be used to simulate various crime prevention strategies and their potential impact on crime rates.
7. Social Network Analysis: Crime often spreads through social networks, as criminals influence each other's behavior and share information about potential targets. Social network analysis techniques can be used to model these networks and predict how changes in social relationships may affect crime rates.
8. Ensemble Methods: Ensemble methods combine the predictions of multiple models to improve overall accuracy. Techniques like model averaging, stacking, and boosting can be used to create ensemble models that outperform individual algorithms.
9. Text Analysis: Analysis of textual data, such as police reports, social media posts, and news articles, can provide valuable insights into crime patterns and trends. Natural language processing (NLP) techniques can be used to extract relevant information and incorporate it into predictive models.
10. Evaluation and Validation: Regardless of the methodology used, it's crucial to rigorously evaluate and validate predictive models to ensure their accuracy and reliability. Techniques like cross-validation, holdout validation, and validation against historical data can help assess model performance and generalizability.

By combining these methodologies and continually refining predictive models with new data and insights, researchers and law enforcement agencies can improve their ability to forecast and prevent crime effectively.

Key Frame Selection:

Key frame selection in crime rate prediction refers to the process of identifying the most informative and representative time periods or intervals within historical data for modeling and forecasting future crime rates. Here's how key frame selection can be approached in crime rate prediction:

1. Identifying Significant Events: Key frames often correspond to periods marked by significant events or changes in the underlying factors influencing crime rates, such as economic downturns, policy changes, natural disasters, or spikes in criminal activity. Identifying these events and selecting corresponding time periods as key frames can help capture their impact on crime rates.
2. Temporal Segmentation: The historical crime data can be segmented into distinct temporal intervals based on various criteria, such as seasons, years, months, weeks, or specific events. Each segment can be evaluated for its relevance and significance in understanding crime trends and predicting future rates.
3. Trend Analysis: Analyzing the temporal trends within the crime data can help identify periods characterized by consistent patterns or deviations from the norm. Key frames may correspond to periods of significant increase or decrease in crime rates, inflection points, or cyclical patterns.
4. Seasonal Decomposition: Decomposing the time series data into its seasonal, trend, and residual components can help identify key frames corresponding to anomalous or impactful variations in crime rates. These components can be analyzed separately to determine their contribution to overall crime trends.
5. Anomaly Detection: Key frame selection can also be guided by detecting anomalous or outlier time periods within the historical data. These anomalies may represent unique events or circumstances that warrant special attention in the predictive modeling process.
6. Domain Expertise: In addition to quantitative analysis, insights from domain experts, such as criminologists, law enforcement officials, or urban planners, can inform key frame selection by providing context about the socio-economic, demographic, and environmental factors influencing crime rates during specific time periods.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, May 2024

7. Evaluation Metrics: Key frames can be evaluated based on their predictive performance in forecasting future crime rates. Metrics such as forecast accuracy, stability, and generalizability can guide the selection of key frames that yield the most reliable and actionable predictions.
8. Dynamic Selection: Key frame selection may be an iterative process, where the choice of frames evolves over time as new data becomes available and the predictive model is refined. Continuously monitoring model performance and updating key frames accordingly ensures that the predictive model remains relevant and effective.

By carefully selecting key frames that capture the most salient features and dynamics of historical crime data, predictive models can be trained to better anticipate future crime rates and inform targeted intervention strategies.

IV. PROPOSED SYSTEMS

Machine Learning Algorithms applied

A. Nearest Neighbour

K-Nearest Neighbour is one of the simplest Machine Learning algorithms based on Supervised Learning technique.

K-NN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories.

K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

K-NN algorithm can be used for Regression as well as for Classification but mostly it is used for the Classification problems.

K-NN is a non-parametric algorithm, which means it does not make any assumption on underlying data.

It is also called a lazy learner algorithm because it does not learn from the training set immediately instead it stores the dataset and at the time of classification, it performs an action on the dataset.

KNN algorithm at the training phase just stores the dataset and when it gets new data, then it classifies that data into a category that is much similar to the new data.

Example: Suppose, we have an image of a creature that looks similar to cat and dog, but we want to know either it is a cat or dog. So for this identification, we can use the KNN algorithm, as it works on a similarity measure. Our KNN model will find the similar features of the new data

2. SVM

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called support vectors, and hence the algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified

Example: SVM can be understood with the example that we have used in the KNN classifier. Suppose we see a strange cat that also has some features of dogs, so if we want a model that can accurately identify whether it is a cat or dog, so such a model can be created by using the SVM algorithm. We will first train our model with lots of images of cats and dogs so that it can learn about different features of cats and dogs, and then we test it with this strange creature. So as support vector creates a decision boundary between these two data (cat and dog) and choose extusing a decision boundary or hyperplane:

3. Decision Tree

Decision Tree is a Supervised learning technique that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where internal nodes represent the features of a dataset, branches represent the decision rules and each leaf node represents the outcome.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 4, May 2024

In a Decision tree, there are two nodes, which are the Decision Node and Leaf Node. Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed on the basis of features of the given dataset.

It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.

It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.

In order to build a tree, we use the CART algorithm, which stands for Classification and Regression Tree algorithm.

A decision tree simply asks a question, and based on the answer (Yes/No), it further split the tree into subtrees.

V. IMPLEMENTATION

The implementation details include the machine learning approach.

Data-collection: The data collection for the implementation is from the Kaggle. The dataset is freely available. The record collected is almost 63000.

Pre-processing: Once the dataset is collected, it must be pre-processed to get the clean dataset. The pandas and NumPy libraries are available in python for the pre-processing. it is removing of empty values from the dataset or repeated records should be removed.

Analysis: The analysis includes the graphical representation of different values to analyse the dataset property. The different graphs are plotted by Matplotlib libraries. The graphical analysis gives a direction towards the prediction.

Training and Testing: The dataset is divided into training and testing. Generally, 70 % dataset is kept for training and 30% for testing. The dataset ratio can be 70: 30 or 80:20. Validation: Once the model is created, it should be validated with the real-time data valu.

VI. OBJECTIVES OF THE PROJECT

The main objective of the project is to predict the crime rate and analyze the crime rate to be happened in future. Based on this Information the officials can take charge and try to reduce the crime rate. →The concept of Multi Linear Regression is used for predicting the graph between the Types of Crimes (Independent Variable) and the Year (Dependent Variable) →The system will look at how to convert crime information into a regression problem, so that it will help detectives in solving crimes faster. →Crime analysis based on available information to extract crime patterns. Using various multi linear regression techniques, frequency of occurring crime can be predicted based on territorial distribution of existing data and Crime recognition.

VII. CONCLUSION

We can see the Prediction, analysis and Visualisation By measuring the accuracy of the different algorithms, we found that the most suitable algorithm for predicting the crime is based on various Machine Learning algorithms. The algorithm will be a great asset for police, security and intelligence since it is trained on a huge collection of historical data and has been chosen after being tested on a sample data.The project demonstrates the machine learning model to predict the crime prediction with more accuracy as compared to previously implemented machine learning models.

REFERENCES

[1]. Chen, Ling, and Xu Lai. "Comparison between ARIMA and ANN models used in short-term wind speed forecasting." Power and Energy Engineering Conference (APPEEC), 2011 Asia- Pacific. IEEE, 2011.

[2]. Agarwal, Jyoti, Renuka Nagpal, and Rajni Sehgal. "Crime analysis using K-means clustering." International Journal of Computer Applications 83.4 (2013).

[3]. Sathyadevan, Shiju, and Surya Gangadharan. "Crime analysis and prediction using data mining." Networks & Soft Computing (ICNSC), 2014 First International Conference on. IEEE, 2014

[4]. McClendon, Lawrence, and Natarajan Meghanathan. "Using machine learning algorithms to analyse crime data." Machine Learning and Applications: An International Journal (MLAIJ) 2.1 (2015).

[5]. Kansara, Chirag, et al. "Crime mitigation at Twitter using Big Data analytics and risk modelling." Recent Advances and Innovations in Engineering (ICRAIE), 2016 International Conference on. IEEE, 2016.